

Manuscript version: Author's Accepted Manuscript

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

Persistent WRAP URL:

<http://wrap.warwick.ac.uk/124273>

How to cite:

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

Copyright and reuse:

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

Publisher's statement:

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk.

Monitoring Networked Infrastructure with Minimum Data via Sequential Graph Fourier Transforms

Zhuangkun Wei¹, Alessio Pagani², Weisi Guo^{1,2*}

¹ School of Engineering, University of Warwick, Coventry, United Kingdom

² Alan Turing Institute, London, United Kingdom

Abstract—Many urban infrastructures contain complex dynamics embedded in spatial networks. Monitoring using Internet-of-Things (IoT) sensors is essential for ensuring safe operations. An open challenge is given an existing sensor network, where best to collect the minimum amount of representative data. Here, we consider an urban underground water distribution network (WDN) and the problem of contamination detection. Existing topology-based approaches link complex network (e.g. Laplacian spectra) to optimal sensing selections, but neglects the underpinning fluid dynamics. Alternative data-driven approaches such as compressed sensing (CS) offer limited data reduction.

In this work, we introduce a principal component analysis based Graph Fourier Transform (PCA-GFT) method, which can recover the full networked signal from a dynamic subset of sensors. Specifically, at each time step, we are able to predict which sensors are needed for the next time step. We do so, by exploiting the spatial-time correlations of the WDN dynamics, as well as predicting the sensor set needed using sparse coefficients in the transformed domain. As such, we are able to significantly reduce the number of samples compared with CS approaches. The drawback lies in the computational complexity of a data collection point (DCP) updating the PCA-GFT operator at each time-step. The experimental results show that, on average, with nearly 40% of the sensors reported, the proposed PCA-GFT method is able to fully recover the networked dynamics.

Index Terms—infrastructure monitoring, network dynamics, complex networks, graph sampling, water distribution network

I. INTRODUCTION

Urban infrastructure monitoring is challenging when there are networked dynamics, causing cascade effects. Examples include traffic jams, electricity outages, and contamination in the water supply. Cascade effects are caused by both the coupling dynamic between junctions/nodes, as well as the overall topology of the network (e.g. multi-scale feedback loops) [1], [2]. As such it is critical to monitor networked infrastructures using sensors. Indeed, this is a critical part of the wider Digital Twin initiative [3]. However, given a sensor network, it is undesirable for every sensor to transmit data all the time, which can lead to poor sensor battery life and low radio spectral efficiency.

Here, we consider an underground urban water distribution network (WDN), where transmitting digital data is challenging yet essential [4]. In particular, we focus on the threat of contamination [5] from a variety of contamination run-off events [6] (see Fig. 1). To monitor this, installations of an Internet-of-Things (IoT) monitoring sensor in each junction have been studied. One main challenge lies in how to extend the life-span of the sensors, due to the difficulty in repairing and recharging

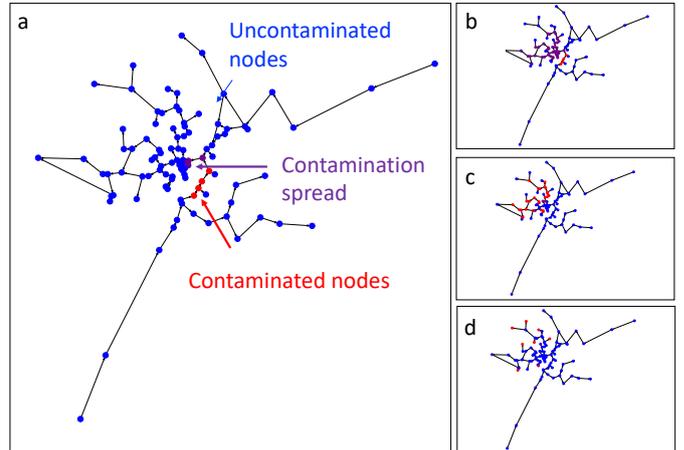


Fig. 1. Simulation of contamination in an urban WDN: (a) initial contamination nodes (red), subsequent spread nodes (purple), and uncontaminated nodes (blue); and (b-d) spread process at later points in time.

and sensors underground. This raises the necessity of how to select minimum number of sensors to detect and report their data to a data collection point (DCP), in which total information can be recovered without hindering the detection accuracy.

A. Literature Review

Aside from human engineering experience based sensor selection, there are two main approaches to optimising sensor selection.

The first group resorts to the graph spectrum analysis [7]–[10], aiming to identify the most influential points on the base of the topological structure of the networks (e.g. via the Laplacian operator [11]–[13]). However, these approaches do not consider the underlying fluid dynamics and assume that the topology dominates. As such, it is important to create an approach that considers both the complex network topology and the contamination signals. Indeed, works that map network topology with explicit dynamics have been progressed in [1] for 1-dimensional ordinary differential equation (ODE) dynamics. However, the challenge with WDNs is that the underlying Navier-Stokes partial differential equation (PDE) dynamics with dynamic Reynolds numbers is high dimensional and difficult to approximate using mean field theory [1]. As such, an analysis of the optimal sampling points as

a function of both the network topology and the dynamic equations is not possible.

The second group focuses on the data-structure instead of the network topology, including the data-driven graph sampling method [14], and the compressed sensing (CS) schemes [15]–[18]. In our previous study [14], a data-driven graph sampling method has been proposed, premised on a prior knowledge of the networked data. However, this method may malfunction when such prior knowledge cannot characterize the real signal. In [15]–[17], three CS approaches have been proposed, aiming to recover the equivalently transformed sparse-representation of the WDN signal. However, their used Discrete Cosine Transform (DCT) matrix [18] is only useful when the networked signals are closely correlated (e.g., the pressure data in one pipe). When it comes to address the data that are less correlated (e.g., the contaminant data indexed on different junctions), the DCT matrix cannot sparsely represent such networked data, which leads to the infeasibility of the CS framework (illustrated in Fig. 4). Indeed, the sequential CS that adapts the transformation matrix via the principal component analysis (PCA) has been proposed in [18], which is capable of sparsely representing the contaminant data in WDN. However, the unknown positions (subscripts) of the sparse coefficients limits its ability to reduce the number of samples (which is explained by the Theorem 1, and via Fig. 4).

B. Contributions and Organization

In this work, we propose a principal component analysis based Graph Fourier Transform (PCA-GFT) method in order to sampling and recovering the networked dynamical signals in WDNs. This approach belongs to the broad family of graph signal processing [9], [10], which states that, if a vector signal of size $N \times 1$ has $\gamma < N$ nonzero coefficients when transformed by a matrix \mathbf{F}^{-1} , then only a subset of nodes can be sampled for full signal recovery. In the context of the time-varying networked signals, the challenge is converted to how to find such \mathbf{F}^{-1} by using only the previous recovered signals. Compared with the PCA-CS in [18], and the adaptive CS, the proposed PCA-GFT method can predict the subscripts of the nonzero coefficients of the transformed current signal, therefore can reduce the number of reported sensors [9]. In comparison with our previously proposed data-driven approach in [14], the new PCA-GFT method avoids the usages of the prior knowledge of the networked data, making it possible to deal with the case in which the prior knowledge cannot characterize the real signals.

The rest of paper is structured as follows. In Section II, we describe the nonlinear dynamical WDN system model, and the aim of this paper. In Section III, we elaborate the proposed PCA-GFT method. In Section IV, the distinction between the new proposed PCA-GFT, the previously proposed data-driven approach in [14], and the PCA-CS in [18] is clarified. Section V illustrates the recovery performance. We finally conclude the paper in Section VI.

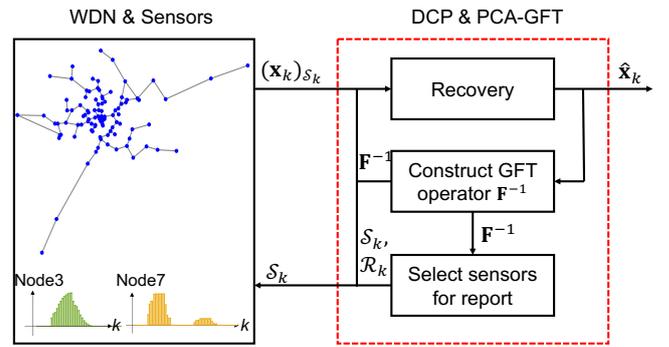


Fig. 2. Illustration of the urban water distribution network (WDN), and the proposed principal component analysis -based Graph Fourier Transform (PCA-GFT) method, which aims to sample and recover the networked signals. The selected monitoring sensors on the nodes (junctions, reservoirs, or tanks) report their data $(\mathbf{x}_k)_{S_k}$ to the data collection point (DCP). The DCP recovers the networked signal as $\hat{\mathbf{x}}_k$ via the collected data and the GFT operator \mathbf{F}^{-1} that is constructed by the previous recovery results.

II. MODEL FORMULATION

The configuration of the WDN can be abstracted as a static graph, denoted as $G(\mathcal{V}, \mathbf{W})$, where $\mathcal{V} = \{1, 2, \dots, N\}$ represents the indices of the total $N \in \mathbb{N}^+$ nodes, and \mathbf{W} gives the weighted adjacency matrix, of which the positive element $w_{n,m} > 0$ accounts for the link from vertex m to vertex n . Here the nodes can be the junctions, the reservoirs, or the tanks, and the links can be pumps or the pipes [19]. On each node of $G(\mathcal{V}, \mathbf{W})$, a sensor is deployed in order to (i) sense the information of the interest, (ii) communicate with the data collection point (DCP). The purpose of the DCP is to (i) recover the time-varying networked signal from the reported data, and (ii) broadcast to sensors whether they should sense and report their data. The illustration of the WDN is provided in Fig. 2.

The information of the interest is the chemical pollutants propagated over the network, which is characterized by a discrete-time matrix data of size $N \times K$, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_K]$. Here, $N = |\mathcal{V}|$ gives the number of vertices of $G(\mathcal{V}, \mathbf{W})$, and $K \in \mathbb{N}^+$ is the total discrete time steps. Hereby, each column \mathbf{x}_k , $1 \leq k \leq K$ contains the data of N nodes of discrete time-step k . Illustrations of the contamination spread process and the contaminant data on two nodes are provided in Fig. 1 and Fig. 2.

As such, the aim of this paper is to recover the real-time signal \mathbf{x}_k via the samples from a sampling vertex set, denoted as $S_k \in \mathcal{V}$.

III. PRINCIPAL COMPONENT ANALYSIS BASED GRAPH SAMPLING

In this section, we elaborate our PCA-GFT method, which aims to sample and recover the networked signal \mathbf{x}_k of WDN without the prior knowledge of the signal. The main idea is to (i) construct a GFT operator that can sparsely represent the current signal, and (ii) adopt the graph sampling theory to select the sensors for report, and recover the data from the

reported samples at DCP. Before we start, we give a brief overview of the graph sampling theory.

Definition 1: [9] A graph signal \mathbf{x} is called $\gamma \in \mathbb{N}$ -support with respect to the GFT operator \mathbf{F}^{-1} , if $\tilde{\mathbf{x}} = \mathbf{F}^{-1}\mathbf{x}$ has only γ non-zero elements with known positions¹.

Definition 2: [9] The set of γ -support graph signals with respect to \mathbf{F}^{-1} in \mathbb{R}^N is a subspace denoted by $BL(\mathcal{R}, \mathbf{F}^{-1})$, if the subscripts of their non-zeros elements are in \mathcal{R} with $\mathcal{R} \subset \mathcal{V}$ and $|\mathcal{R}| = \gamma$.

Theorem 1: [9], [20], [21] For any $\mathbf{x} \in BL(\mathcal{R}, \mathbf{F}^{-1})$, there exists a subset $\mathcal{S} \subset \mathcal{V}$ such that

$$\hat{\mathbf{x}} = \mathbf{F}_{\mathcal{V}\mathcal{R}} \cdot (\mathbf{F}_{\mathcal{S}\mathcal{R}}^T \cdot \mathbf{F}_{\mathcal{S}\mathcal{R}})^{-1} \cdot \mathbf{F}_{\mathcal{S}\mathcal{R}}^T \cdot \mathbf{x}_{\mathcal{S}}.$$

Such \mathcal{S} satisfies

$$\text{rank}(\mathbf{F}_{\mathcal{S}\mathcal{R}}) = |\mathcal{R}|,$$

where $\mathbf{F}_{\mathcal{S}\mathcal{R}}$ ($\mathbf{F}_{\mathcal{V}\mathcal{R}}$) denotes the sub-matrix of \mathbf{F} with rows selected via subscripts in \mathcal{S} (\mathcal{V}) and columns selected via subscripts in \mathcal{R} , and $\mathbf{x}_{\mathcal{S}}$ denotes the sampled vector of \mathbf{x} by selecting subscripts in \mathcal{S} .

Theorem 1 indicates the existence of a minimum number of selected sensors, which equals to the number of nonzeros in $\tilde{\mathbf{x}} = \mathbf{F}^{-1} \cdot \mathbf{x}$ (i.e., $|\mathcal{S}| = |\mathcal{R}| = \gamma$). By contrast, the compressed sensing based algorithms require at least $O(\gamma \log(N/\gamma))$ sensors for full recovery. This advantage of reduction is due to (i) the design of their GFT operator \mathbf{F}^{-1} that ensures the γ -support, and (ii) more importantly, the known subscripts of the nonzeros of $\tilde{\mathbf{x}} = \mathbf{F}^{-1} \cdot \mathbf{x}$.

A. PCA based GFT Operator

Given the concept of graph sampling theory, the GFT operator \mathbf{F}^{-1} should ensure two properties. First, \mathbf{x}_k should be γ -support with respect to \mathbf{F}^{-1} . Second, the subscripts of the nonzero coefficients in $\tilde{\mathbf{x}}_k = \mathbf{F}^{-1}\mathbf{x}_k$ should be predictable via the previous recovery results. These two constitute the reason to resort the PCA technique, as it has the ability to sparsely represent a networked signal [18].

1) *Construction of PCA-GFT operator:* For each time-step k , we construct the GFT operator \mathbf{F}^{-1} via the previous recoveries. Let denote the recovery as $\hat{\mathbf{x}}_{k-L}, \dots, \hat{\mathbf{x}}_{k-1}$. The mean and the covariance matrix are computed respectively as:

$$\bar{\mathbf{x}} = \frac{1}{L} \sum_{l=1}^L \hat{\mathbf{x}}_{k-l}, \quad (1)$$

$$\Sigma = \frac{1}{L} \sum_{l=1}^L (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}) \cdot (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}})^T, \quad (2)$$

where L is the lag accounting for the correlations, i.e., $\mathbf{x}_k = f(\mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-L})$. The effect of L on recovery performance

¹Different from the concept of γ -sparse vector in compressed sensing where the positions of the non-zero elements are unknown, in the case of graph sampling theory, we know the positions (subscripts) of these non-zero elements. Hence, the size of the selected sensor set $|\mathcal{S}|$ can be smaller as opposed to that used by CS, which is explained after Theorem 1

is analyzed in Section V. Then, the qr-factorization is used to derive the GFT operator, i.e.,

$$\Sigma = \mathbf{Q} \cdot \mathbf{R}, \quad (3)$$

$$\mathbf{F}^{-1} = \mathbf{Q}^{-1}. \quad (4)$$

2) *Sparsity Analysis:* After the derivation of the GFT operator \mathbf{F}^{-1} in Eqs. (1)-(4), we here prove that \mathbf{F}^{-1} satisfies two properties (i.e., the $\gamma < N$ -support property of \mathbf{x}_k , and predictability of nonzeros' subscripts of $\tilde{\mathbf{x}}_k$).

Given the spatial-time correlation of the signal, we express the current signal \mathbf{x}_k as:

$$\mathbf{x}_k = \sum_{l=1}^L c_l \cdot \hat{\mathbf{x}}_{k-l} + \mathbf{r}_k, \quad (5)$$

where c_l represents the corresponding coefficients, and \mathbf{r}_k denotes the residue component. By subtracting $\bar{\mathbf{x}}$ on both sides, Eq. (5) is computed as:

$$\mathbf{x}_k - \bar{\mathbf{x}} = \sum_{l=1}^L c_l \cdot (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}) + \left(\sum_{l=1}^L c_l - 1 \right) \bar{\mathbf{x}} + \mathbf{r}_k \quad (6)$$

According to Eq. (2), $\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}$ can be computed via the linear combination of the columns of Σ_k . As such, we re-write $\mathbf{x}_k - \bar{\mathbf{x}}$ as:

$$\mathbf{x}_k - \bar{\mathbf{x}} = \Sigma_k \cdot \boldsymbol{\beta} + \eta \cdot \bar{\mathbf{x}} + \mathbf{r}_k, \quad (7)$$

where $\boldsymbol{\beta} = [\beta_1, \dots, \beta_N]^T$ gives the coefficient vector, and $\eta = \sum_{l=1}^L c_l - 1$. The transformation of $\mathbf{x}_k - \bar{\mathbf{x}}$ with respect to \mathbf{F}^{-1} is

$$\begin{aligned} \tilde{\mathbf{x}}_k - \tilde{\bar{\mathbf{x}}} &= \mathbf{F}^{-1} \cdot (\mathbf{x}_k - \bar{\mathbf{x}}) \\ &= \mathbf{Q}^{-1} \cdot \Sigma_k \cdot \boldsymbol{\beta} + \eta \cdot \mathbf{Q}^{-1} \cdot \bar{\mathbf{x}} + \mathbf{Q}^{-1} \cdot \mathbf{r}_k \\ &= \underbrace{\mathbf{R} \cdot \boldsymbol{\beta} + \eta \cdot \mathbf{Q}^{-1} \cdot \bar{\mathbf{x}}}_{\text{predictable subscripts of nonzeros}} + \mathbf{Q}^{-1} \cdot \mathbf{r}_k. \end{aligned} \quad (8)$$

In Eq. (8), it is intuitive that the subscripts of the non-zeros in $\eta \cdot \mathbf{Q}^{-1} \cdot \bar{\mathbf{x}}$ can be obtained via direct computation. Then, the vector $\mathbf{R} \cdot \boldsymbol{\beta}$ has at most the first ($\gamma \leq L$)-row of non-zero elements. This is because \mathbf{R} is an upper-triangular matrix with $\text{rank } \gamma = \text{rank}(\mathbf{R}) = \text{rank}(\Sigma) = \text{rank}(\sum_{l=1}^L (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}) \cdot (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}})^T) \leq \sum_{l=1}^L \text{rank}((\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}) \cdot (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}})^T) = L$.

These two indicate the sparsity (i.e., $\gamma \leq L$) and the predictability of nonzeros' subscripts of $\tilde{\mathbf{x}}_k$. Further, let denote the set of such nonzeros' subscripts as $\mathcal{R}_{k|k-1}$. According to Eq. (8), $\mathcal{R}_{k|k-1}$ can be predicted via the transformed signals of $\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}$, and the nonzeros' subscripts of $\mathbf{Q}^{-1} \cdot \bar{\mathbf{x}}$, i.e.,

$$\begin{aligned} \mathcal{R}_{k|k-1} &= \left\{ n \mid \bigcup_{l=1}^L (\mathbf{F}^{-1} \cdot (\hat{\mathbf{x}}_{k-l} - \bar{\mathbf{x}}))_n \neq 0 \cup (\mathbf{F}^{-1} \cdot \bar{\mathbf{x}})_n \neq 0 \right\}, \end{aligned} \quad (9)$$

where $(\cdot)_n$ denotes the n th element. According to Theorem 1, this $\mathcal{R}_{k|k-1}$ maps to a set of selected sensors $\mathcal{S}_{k|k-1}$, which can be used to recover the component of $\mathbf{x}_k - \bar{\mathbf{x}}$, i.e., $\Sigma\boldsymbol{\beta} + \eta\bar{\mathbf{x}}$ in Eq. (7). For the residual component, i.e., \mathbf{r}_k in Eq. (7), an extra set of sensors, denoted as \mathcal{S}^* will be selected for monitoring. We next study how to determine $\mathcal{S}_{k|k-1}$ and \mathcal{S}^* .

B. Selection of Sampling Node Set

Once we derive the GFT operator \mathbf{F}^{-1} from Eqs. (1)-(4), we design the selection of the sensors for report. Let denote \mathcal{S}_k as the sensor set for report that ensures the full recovery. In accordance with Eq. (8), \mathcal{S}_k should consist of two subsets, i.e.,

$$\mathcal{S}_k = \mathcal{S}_{k|k-1} \cup \mathcal{S}^*, \quad (10)$$

where $\mathcal{S}_{k|k-1}$ is the predicted sensor set that is derived from $\mathcal{R}_{k|k-1}$, and \mathcal{S}^* accounts for the extra sensor set for monitoring the residue component \mathbf{r}_k in Eq. (5).

The determination of $\mathcal{S}_{k|k-1}$ is given by Theorem 1, i.e.,

$$\text{rank}(\mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}}) = |\mathcal{R}_{k|k-1}|. \quad (11)$$

Given $\mathcal{R}_{k|k-1}$ from Eq. (9), we compute $\mathcal{S}_{k|k-1}$ by finding the $|\mathcal{R}_{k|k-1}|$ smallest singulars of $\mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}}$, i.e.,

$$\mathcal{S}_{k|k-1} = \underset{\mathcal{S}_{k|k-1} \subset \mathcal{V}}{\text{argmax}} \sigma_{\min}(\mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}}), \quad (12)$$

where $\sigma_{\min}(\cdot)$ denotes the minimum singular of the matrix. Based on Eq. (12), the recursive *greedy* algorithm can be implemented by finding and adding the row, i.e., $\mathcal{S}_{k|k-1} \leftarrow \mathcal{S}_{k|k-1} \cup \{n\}$, such that $n = \underset{i}{\text{argmax}} \sigma_{\min}(\mathbf{F}_{(\mathcal{S}+\{i\})\mathcal{R}})$.

The design of \mathcal{S}^* aims to predict the potential nodes where the pollutant is burst (i.e., the node n with $(\mathbf{x}_{k-1})_n = 0$, but $(\mathbf{x}_k)_n \neq 0$). As such, we rely on the topology of $G(\mathcal{V}, \mathbf{W})$, by estimating a rough outcome via the multiplication of adjacency matrix and the previous recovery, i.e.,

$$\mathcal{S}^* \subset \left\{ n \mid (\mathbf{W} \cdot \hat{\mathbf{x}}_{k-1})_n \neq 0 \cap (\hat{\mathbf{x}}_{k-1})_n = 0 \right\}. \quad (13)$$

C. Signal Recovery

With the help of the construction of the sampling node set \mathcal{S}_k in Eqs. (10)-(13), the sensors with indices belonging to \mathcal{S}_k can report their data to the DCP, which then collects the samples as $(\mathbf{x}_k)_{\mathcal{S}_k}$. The recovery process of the DCP can be divided into two parts. The first part is referred to the Theorem 1, i.e.,

$$\hat{\mathbf{x}}_k = \mathbf{F}_{\mathcal{V}\mathcal{R}_{k|k-1}} \cdot \left(\mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}}^T \cdot \mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}} \right)^{-1} \cdot \mathbf{F}_{\mathcal{S}_{k|k-1}\mathcal{R}_{k|k-1}}^T \cdot ((\mathbf{x}_k)_{\mathcal{S}_{k|k-1}} - \bar{\mathbf{x}}_{\mathcal{S}_{k|k-1}}) + \bar{\mathbf{x}}. \quad (14)$$

Then, for the second part that relies on \mathcal{S}^* , we replace the corresponding elements in $\hat{\mathbf{x}}_k$ with $(\mathbf{x}_k)_{\mathcal{S}^*}$, i.e.,

$$(\hat{\mathbf{x}}_k)_{\mathcal{S}^*} = (\mathbf{x}_k)_{\mathcal{S}^*}. \quad (15)$$

IV. DISTINGUISH WITH TWO DATA-DRIVEN SCHEMES

In this section, we distinguish our proposed PCA-GFT method with other two data-driven sampling schemes. The first one is provided in our previous work [14]. The second is the widely-used PCA-CS in [18].

A. Data-driven Static Graph Sampling

In our previous work [14], a data-driven graph sampling algorithm has been proposed leveraged on a prior knowledge of the data, denoted as $\bar{\mathbf{X}}$ (which can be derived via the simulators of the designed WDN). We assume (i) the real data \mathbf{X} has limited deviation from $\bar{\mathbf{X}}$, i.e., $\|\mathbf{X} - \bar{\mathbf{X}}\|_2 < \epsilon$ for a small ϵ , and (ii) $\text{rank}(\mathbf{X}) = \gamma < N$. In this way, in order to make $\bar{\mathbf{X}}$ being $\gamma < N$ -support with respect to the GFT operator $\bar{\mathbf{F}}^{-1}$, we designed the GFT operator based on the qr-factorization of the maximally linearly independent columns of $\bar{\mathbf{X}}$ (denoted as $\bar{\mathbf{X}}_{\text{m.i.c.}}$), i.e.,

$$\bar{\mathbf{X}}_{\text{m.i.c.}} = \bar{\mathbf{F}} \cdot \bar{\mathbf{R}}. \quad (16)$$

Also, given the property of qr-factorization, the non-zero elements of $\bar{\mathbf{X}} = \bar{\mathbf{F}}^{-1} \cdot \mathbf{X}$ will be located at first r rows, and thus $\mathcal{R} = \{1, \dots, \gamma\}$. With the derivations of the GFT operator and \mathcal{R} , we utilized Theorem 1 to recover the data.

Compared with our previous data-driven graph sampling, the proposed PCA-GFT does not rely on the prior knowledge of the data, i.e., $\bar{\mathbf{X}}$, which if not reliable, may result in difficulties for monitoring applications. Then, it is noteworthy that the proposed PCA-GFT requires more energy expenditure for the DCP end, as it has to update the GFT operator and broadcasts the selected sensor set \mathcal{S}_k at each time-step k .

B. PCA-CS

PCA-CS has been proposed in [18], which aims at sampling and recovering the networked signals via the subset of sensors². The PCA-CS uses the eigen-vector matrix Ψ as the transformation matrix, i.e.,

$$\mathbf{x}_k - \bar{\mathbf{x}} = \Psi \cdot \boldsymbol{\alpha}_k, \text{ with } \Sigma = \Psi \cdot \Lambda \cdot \Psi^{-1} \quad (17)$$

where $\boldsymbol{\alpha}_k$ denotes the sparse coefficients. $\bar{\mathbf{x}}$ and Σ are given from Eqs. (1)-(2). Then, the sampling and recovery challenge can be pursued by selecting $\mathcal{S}_k \subset \mathcal{V}$ such that the restricted isometric property (RIP) is satisfied, i.e.,

$$1 - \delta_{2\gamma} \leq \frac{\|\Psi_{\mathcal{S}_k\mathcal{V}} \cdot \boldsymbol{\alpha}\|_2^2}{\|\boldsymbol{\alpha}_k\|_2^2} \leq 1 + \delta_{2\gamma}, \quad \gamma = \|\boldsymbol{\alpha}_k\|_0 \quad (18)$$

for any 2γ -sparse $\boldsymbol{\alpha}$ and some $\delta_{2\gamma} \in [0, 1]$, where $\|\cdot\|_2$ and $\|\cdot\|_0$ represent 2-norm and 0-norm respectively. Then, given the samples of k time-step, $\boldsymbol{\alpha}_k$ can be recovered via the convex optimization, or the orthogonal matching pursuit (OMP), and thus $\hat{\mathbf{x}}_k = \Psi \cdot \hat{\boldsymbol{\alpha}}_k + \bar{\mathbf{x}}$ can be computed.

The major difference between the proposed PCA-GFT and the PCA-CS is whether the positions of the sparse coefficients can be predicted. In the proposed PCA-GFT, it is proved in Eq. (8) that the subscripts of the non-zero elements of $\hat{\mathbf{x}}_k = \mathbf{F}^{-1} \cdot \mathbf{x}_k$ can be predicted from that of $\hat{\mathbf{x}}_{k-l} = \mathbf{F}^{-1} \cdot \hat{\mathbf{x}}_{k-l}$. With the information of the non-zero positions, the PCA-GFT method can reduce the size of selected sensor set to near γ (validated by Theorem 1), as opposed to the PCA-CS which requires $O(\gamma \log(N/\gamma))$. This will be further analyzed in Section V.

²Here, similar to the proposed PCA-GFT, the selection of sensors of PCA-CS also changes with the time

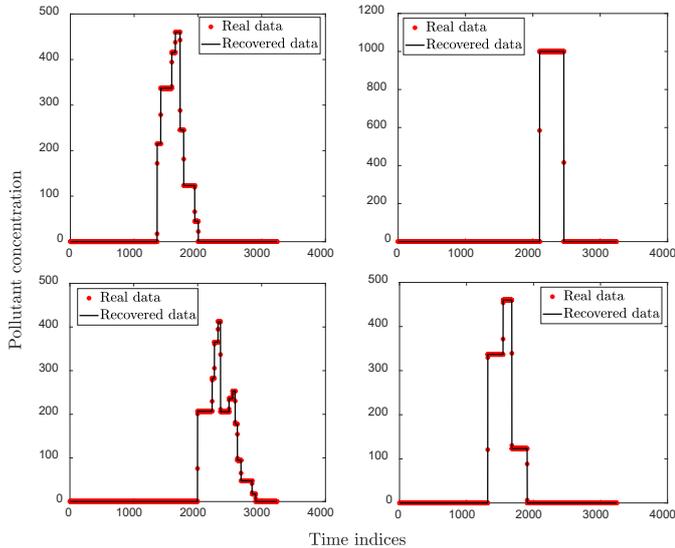


Fig. 3. Illustration of 4 examples of real and recovered data of the proposed PCA-GFT method.

V. RESULTS

In the following analysis, the recovery performance versus the averaging number of reported sensors of our proposed PCA-GFT method will be evaluated. The recovery accuracy is measured in terms of the mean absolute error (MAE) of the recovered data, i.e.,

$$\text{MAE} = \frac{1}{NK} \sum_{k=1}^K \|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_1, \quad (19)$$

where $\|\cdot\|_1$ denotes the 1-norm. The averaging number of reported sensors is computed as:

$$\overline{|\mathcal{S}|} = \frac{1}{K} \sum_{k=1}^K |\mathcal{S}_k|. \quad (20)$$

The simulations in this work are pursued using the Python package Water Network Tool for Resilience (WNTR) based on EPANET2 [19]. The simulations are executed on Microsoft Azure [22]. The WDN network is configured as $N = 102$ nodes (see Fig. 2), including 100 junctions and 2 reservoirs. For each junction, a random and unknown water-demand is used. The links are pipes with unknown pressures. We simulate 100 different time-varying chemical contaminant propagated over the WDN. Each data is simulated for 3 hours with $K = 3240$ time steps.

A. One Illustration of Recovery Performance

We firstly provide one illustration of our proposed PCA-GFT method in Fig. 3, which presents the comparisons between real data and the recovered data on 4 nodes. We figure out that the perfect recovery is achieved, with averaging number of reported sensors as $\overline{|\mathcal{S}|} = 46 < N = 102$, which is lesser than the half of the total number of sensors.

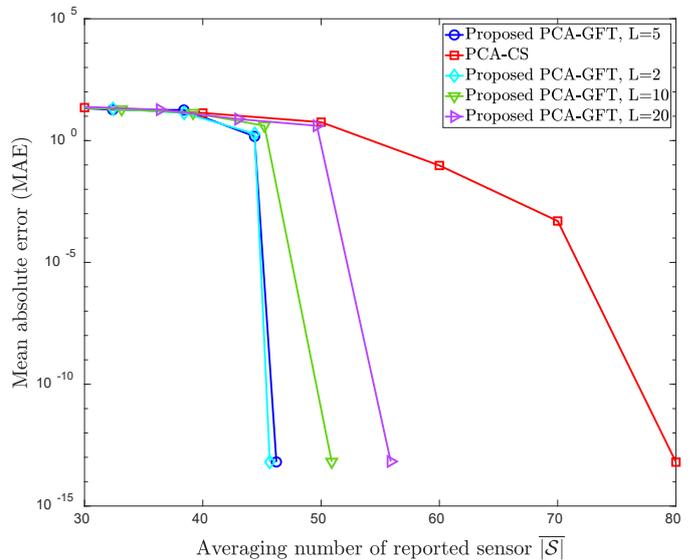


Fig. 4. Comparison of recovery accuracy between proposed PCA-GFT and PCA-CS.

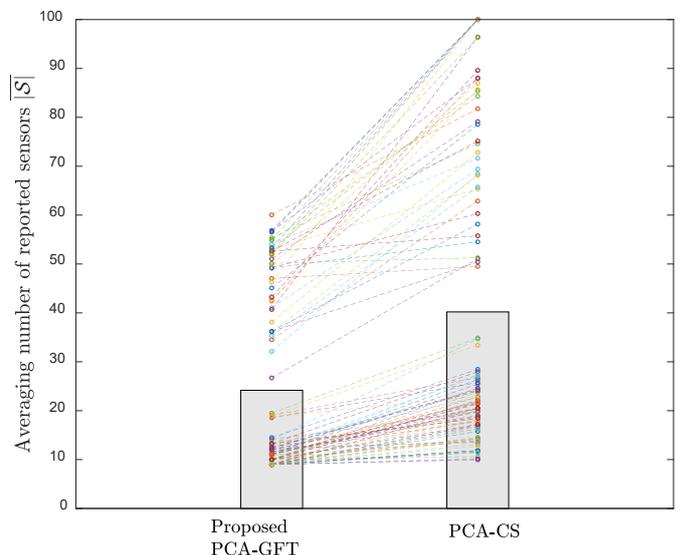


Fig. 5. Comparison of minimum number of reported sensors for full recovery within 100 sets of data.

B. Performance Comparisons

The performance comparison between our proposed PCA-GFT method, and the PCA-CS [18] is illustrated in Figs. 4-5. Fig. 4 presents recovery performance (i.e., MAE) of the two schemes with respect to the changes of the averaging number of reported sensors $\overline{|\mathcal{S}|}$. It is firstly seen that the recovery accuracy (i.e., the MAE) decreases as the length of lag L increases. The reason is give as follows. In Eq. (5), L accounts for (i) the the correlations between the current signal \mathbf{x}_k and the previous states, i.e., $\mathbf{x}_k = f(\mathbf{x}_{k-1}, \dots, \mathbf{x}_{k-L})$, and (ii) the number of sensors in $\mathcal{S}_{k|k-1}$ (i.e., $|\mathcal{S}_{k|k-1}| = \mathcal{R}_{k|k-1} = L$ seen in Eq. (8)). In the context of the WDN where the current contaminant directly evolves from its last state. i.e.,

$\mathbf{x}_k \leftarrow \mathbf{x}_{k-1}$, we should keep a smaller L as $L = 2$ in order to (i) holds the correlations³ and (ii) minimize $|\mathcal{S}_{k|k-1}|$. Hence, $L = 2$ provides the minimum $|\overline{\mathcal{S}}|$ as is illustrated in Fig. 4.

Then, we can observe that the proposed PCA-GFT outperforms the PCA-CS. The former requires approximately $|\overline{\mathcal{S}}| = 46$ sensors for the full recovery, whilst the PCA-CS needs $|\overline{\mathcal{S}}| = 80$ sensors. This can be further demonstrated in Fig. 5, where the minimum $|\overline{\mathcal{S}}|$ is recorded for the full recovery within 100 different data. We can see that the number of reported sensor $|\overline{\mathcal{S}}|$ in our PCA-GFT keeps smaller as opposed to the PCA-CS. The advantage of the sensor reduction provided by the proposed PCA-GFT is attributed to the predictability of the positions (subscripts) of the γ -sparse coefficients of $\tilde{\mathbf{x}}_k = \mathbf{F}^{-1} \cdot \mathbf{x}_k$, ($\gamma = \|\tilde{\mathbf{x}}_k\|_0$). As mentioned in Theorem 1, the knowledge of such subscripts enables the selection of reported node set whose size equals to γ , i.e., $|\mathcal{S}_k| = \gamma$. By contrast, the PCA-CS needs at least $O(\gamma \log(N/\gamma))$ in order to ensure the RIP.

VI. CONCLUSIONS AND DISCUSSION

In order to monitor the networked dynamics on critical urban infrastructures, we proposed a principal component analysis based Graph Fourier Transform (PCA-GFT) method, which can recover the full networked signal from a subset of sensors. The constructed PCA-GFT operator can ensure the sparse property of the networked signal, as well as predicting the sensor set needed by analyzing the previous recovery in the transformed domain. As such, the PCA-GFT is capable of reducing the number of samples compared with the compressed sensing (CS) approaches. The drawback lies in the computational complexity of a data collection point (DCP) updating the PCA-GFT operator at each time-step. The experimental results demonstrate the averaging 40% of the sensors are needed to ensure the full recovery of the networked dynamics. The performance guarantee given by the framework enables us to reduce the number of sampling points in a hierarchical manner whilst loosing accuracy.

The framework is useful beyond the application of water distribution networks (WDNs) and can be applied to a variety of infrastructure sensing (e.g. railways [23]) for digital twin modeling. Current deficiencies include the need for the sensors to switch on and off, and as such many sensors are deployed. Future work will focus on how to improve the trade-off between minimum sensor-side processing, edge and cloud computing [24], and big data fusion for intelligent water distribution networks and digital twin modelling.

Acknowledgements: The authors (A.P. & W.G.) acknowledge funding from the Lloyd's Register Foundation's Programme for Data-Centric Engineering at The Alan Turing Institute, and funding from The Alan Turing Institute under the EPSRC grant EP/N510129/1. The author W.G. acknowledge funding from EPSRC under grant EP/R041725/1. The authors acknowledge Microsoft Corporation for providing cloud resources on Microsoft Azure.

³Even if the length of lag $L = 1$ can maintain the correlation, $L = 1$ is trivial in PCA theory [18].

REFERENCES

- [1] J. Gao, B. Barzel, and A. Barabasi, "Universal resilience patterns in complex networks," *Nature*, vol. 530, 2016.
- [2] A. Pagani, G. Mosquera, A. Alturki, S. Johnson, S. Jarvis, A. Wilson, W. Guo, and L. Varga, "Resilience or Robustness: Identifying Topological Vulnerabilities in Rail Networks," *Royal Society Open Science*, vol. 6, 2019.
- [3] F. Tao and M. Zhang, "Digital twin shop-floor: A new shop-floor paradigm towards smart manufacturing," *IEEE Access*, vol. 5, pp. 20 418–20 427, 2017.
- [4] M. M. Mekonnen and A. Y. Hoekstra, "Four billion people facing severe water scarcity," *Science Advances*, vol. 2, no. 2, 2016.
- [5] V. Pye and R. Patrick, "Ground water contamination in the united states," *Science*, vol. 221, no. 4612, pp. 713–718, 1983.
- [6] W. Ritter, "Pesticide contamination of ground water in the united states - a review," *Journal of Environmental Science and Health, Part B*, vol. 25, no. 1, pp. 1–29, 1990.
- [7] A. Di Nardo, C. Giudicianni, R. Greco, M. Herrera, G. Santonastaso, and A. Scala, "Sensor placement in water distribution networks based on spectral algorithms," *13th International Conference on Hydroinformatics (HIC2018)*, 07 2018.
- [8] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Trans. on Signal Processing*, vol. 64, no. 14, pp. 3775–3789, 2016.
- [9] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. on Signal Processing*, vol. 63, no. 24, pp. 6510–6523, 2015.
- [10] A. Sandryhaila and J. Moura, "Big Data Analysis with Signal Processing on Graphs: Representation and processing of massive data sets with irregular structure," *IEEE Signal Processing Magazine*, vol. 31, 2014.
- [11] F. Archetti, A. Candelieri, and D. Soldi, "Network analysis for resilience evaluation in water distribution networks," *Environmental Engineering and Management Journal*, vol. 14, 2015.
- [12] C. Giudicianni, A. Nardo, M. Natale, R. Greco, G. Santonastaso, and A. Scala, "Topological taxonomy of water distribution systems," *Water*, vol. 10, 2018.
- [13] A. Simone, L. Ridolfi, D. Laucelli, L. Berardi, and O. Giustolisi, "Centrality metrics for water distribution networks," *EPiC Series in Engineering*, vol. 3, 2018.
- [14] Z. Wei, A. Pagani, G. Fu, I. Guymier, W. Chen, J. McCann, and W. Guo, "Optimal sampling of water distribution network dynamics using graph fourier transform," *arXiv preprint arXiv:1904.03437*, 2019.
- [15] R. Du, L. Gkatzikisz, C. Fischione, and M. Xiao, "Energy efficient monitoring of water distribution networks via compressive sensing," in *2015 IEEE International Conference on Communications (ICC)*, 2015, pp. 6681–6686.
- [16] L. Xu, X. Qi, Y. Wang, and T. Moscibroda, "Efficient data gathering using compressed sparse functions," in *2013 Proceedings IEEE INFOCOM*, 2013, pp. 310–314.
- [17] S. Kartakis, G. Tzagarakis, and J. McCann, "Adaptive Compressive Sensing in Smart Water Networks," *MDPI 2nd International Ele. Conf. on Sensors and Applications*, vol. 6, 2019.
- [18] G. Quer, R. Masiero, G. Pillonetto, M. Rossi, and M. Zorzi, "Sensing, compression, and recovery for wsn: Sparse signal modeling and monitoring framework," *IEEE Trans. on Wireless Communications*, vol. 11, no. 10, pp. 3447–3461, 2012.
- [19] L. Rossman, "Epanet 2 users manual," *U.S. Environmental Protection Agency, Washington, D.C., EPA/600/R-00/057*, 2000.
- [20] A. Sandryhaila and J. Moura, "Discrete signal processing on graphs: frequency analysis," *IEEE Trans. on Signal Processing*, vol. 62, 2014.
- [21] A. Anis, A. Gadde, and A. Ortega, "Efficient sampling set selection for bandlimited graph signals using graph spectral proxies," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3775–3789, 2016.
- [22] Microsoft Corporation. Get started with azure. [Online]. Available: <https://docs.microsoft.com/en-gb/azure/>
- [23] O. Jo, Y. Kim, and J. Kim, "Internet of things for smart railway: Feasibility and applications," *IEEE Internet of Things Journal*, vol. 5, no. 2, pp. 482–490, April 2018.
- [24] F. H. Bijarbooneh, W. Du, E. C. . Ngai, X. Fu, and J. Liu, "Cloud-assisted data fusion and sensor selection for internet of things," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 257–268, June 2016.